Project 1 - Classification

CSCI 5523 - Introduction to Data Mining
University of Minnesota

Due - November 02, 2022

Instructions and Experiments

Note: Please read the entire project description before you begin.
This project aims to analyze classification algorithms' performance on several synthetic and real-world data sets. Please follow the steps below:

- First, explore the data sets.
- Next, perform a series of experiments in which you should answer a series of questions. For these experiments, you need to run a python Jupyter notebook.
- Compile your answers in the form of a report.

Python Jupyter Notebooks

We recommend installing Jupyter using Anaconda as it will also install other regularly used packages for scientific computing and data science. Some pointers to setup Jupyter notebooks on your system:

- Video link - *https://www.youtube.com/watch?v=MvN7Wdh0Juk*
- Medium Link - *https://medium.com/@neuralnets/beginners-quick-guide-for-handling-issues-launching-jupyter-notebook-for-python-using-anaconda-8be3d57a209b*
- Tutorials Link - *https://www.dataquest.io/blog/jupyter-notebook-tutorial/*, *https://www.youtube.com/watch?v=3C9E2yPBw7s*

Before you Begin

- Visually explore the data sets in the experiments below, and consider the following:
  - ✓ types of attributes
  - ✓ class distribution
  - ✓ which attributes appear to be good predictors, if any
  - ✓ possible correlation between attributes
  - ✓ any special structure that you might observe

  Note: Discussion of this exploration is not required in the report, but this step will help you get ready to answer the questions that follow
- Use precision and recall to measure performance.
- Your goal is to learn everything that you can about the dataset. Answer the questions below as a starting point, but you should dig further. What more can you discover? The goal of this assignment is to give a helping hand for you to find the most interesting and surprising things.

Report and Submission

- Collect output from your experiments. **Submit all Jupyter notebooks (cell displaying output) electronically as a single zipped file using the Project 1 Canvas submit tool**. A submission not adhering to this policy will not be graded, and you will get a zero.
- Write a report addressing the experiment questions. **The report must be submitted in PDF format electronically using Project 1 Gradescope submission.** Your project will be evaluated based only on what you write in the report.
- Your Jupyter notebook should be submitted electronically - we will look at your output if something is ambiguous in your report. Copy and paste the output from the Jupyter notebook into your report only to the limited extent needed to support your answers.

Problem 1 [20 points]

Files for this problem are under the Experiment 1 folder.

Datasets for experimentation: telecom_churn.csv.

Jupyter notebook: Exploratory data analysis.ipynb.

In this experiment, we will do exploratory data analysis to understand the data better. The dataset contains the record of telecom customers along with the label "churn". Churn = "true" signifies that the customer has left the company, and churn = "false" signifies that the customer is still loyal to the company.

Answer the following questions:

1. How many records are there in the dataset?
2. How many features are there? Name each feature and assign it as binary, discrete, or continuous.
3. As a data scientist, your job is to build a model that identifies customers intending to leave your company. To do that, we prepare our data for the machine learning model. We can have the most advanced algorithm, but our results will be poor if our training data is terrible. According to your intuition, which features are irrelevant? Briefly explain your reasoning.
4. Are there any missing values in the data?
5. What are the average, median, maximum, minimum, and standard deviation values for the continuous features?
6. What is the average number of customer service calls a customer makes to the company?
7. In our dataset, data comes from how many states?
8. What's the distribution of the "Churn" feature? Is this feature skewed?
9. What are the customers' highest and lowest "total day charge"? If we sort the dataset in ascending and descending order by "total day charge", what observation can you make regarding the connection between "total day charge" and "churn" rate?
10. What's the average number of customer service calls made by the user who has churned out of the company? Compare and contrast it with the average number of customer service calls made by the user who is still with the company.
11. Compare and contrast the average values of numerical features for churned and non-churned users. As a data scientist, what strategy will you recommend to the company to retain more customers?
12. Assume you have devised a model which states that if "international plan" = 'no', then the customer will not churn (i.e., "churn" = False). Report accuracy, precision, and recall concerning the "churned" class.

13. Calculate P (churn = True | international plan = ‘yes’), P (churn = False | international plan = ‘yes’), P (churn= True | international plan = ‘no’), P (churn = False | international plan = ‘no’). For a customer who has churned, what are the probabilities that the customer has opted/not opted for the international plan? Similarly, given that the customer has not churned, what are the probabilities that the customer has opted/not opted for the international plan?
14. Calculate the probability of customers leaving the company, given that he has not made any customer service call. Compare and contrast it with the customer making 1,2,3,4,5,6,7,8,9 customer service calls. Plot the probability of customers leaving the company as customer service calls increase.
15. Assume you have devised a model which states that if "international plan" = 'yes' and the number of calls to the service center is greater than 3, then the customer will churn (i.e., "churn" = True). Report accuracy, precision, and recall concerning the "churned" class.

Problem 2 [20 points]

Files for this problem are under the Experiment 2 folder.

Datasets for experimentation: telecom_churn.csv.

Jupyter notebook: Decision Trees and kNN.ipynb.

In this experiment, we will apply and visualize decision trees, apply kNN, finetune parameters, and learn about k-fold cross-validation. To visualize the decision tree, we need to install additional packages: Graphviz and pydotplus.

Answer the following questions:

1. Decision tree classifier *sklearn.tree.DecisionTreeClassifier* has the parameter "max_depth", which defines the maximum depth of the tree, and "criterion", which measures the quality of the split. What happens if we don't specify any value for both parameters?
2. For the synthetic dataset, we separate two classes by training a decision tree. What does the boundary look like when we overfit (max_depth ≥ 4) and underfit (max_depth = 1) the decision tree on data? For both cases, paste the decision tree and the decision boundary from the Jupyter notebook output.
3. For Bank Dataset, what are the 5 different age values that the decision tree used to construct the tree? What is the significance of these 5 values?
4. Given a dataset d, with n samples and m continuous features, what does Standard Scaler *sklearn.preprocessing.StandardScaler* do? Given dataset d = [[0, 0], [0, 0], [1, 1], [1, 1]], write down its scaler transformation.
5. In the section Underfitting and Overfitting (Jupyter notebook), we have classified two datasets (a small one and a big one) using two different decision trees to demonstrate underfitting and overfitting. Briefly describe the experiments performed and what you learn from these experiments.

6. In section Imbalanced Class (Jupyter notebook) , we have trained a couple of classifiers on the balanced and unbalanced datasets and evaluated their accuracy on balanced and unbalanced datasets. Furthermore, we have printed the confusion matrix. Briefly describe the experiments performed and what you learn from these experiments. Also, write down the experiments' precision, recall, and f1 score.
7. In the section Adding irrelevant attributes (Jupyter notebook), we have added an irrelevant attribute to the dataset and have trained a decision tree classifier on it. Based on the test set results, what do you think has happened, and can you connect it with the class material? Briefly describe the experiment done and the intuition developed from these experiments.
8. For the customer churn prediction task, we show that the accuracy of the decision tree is 94% when max_depth is set to 5. What happens to accuracy when we leave the value of max_depth at its default value? Explain the rise/fall of accuracy.
9. How many decision trees do we have to construct if we have to search the two-parameter space, max_depth [1-10] and max_features [4-18]? If we consider 10-fold cross-validation with the above scenario, how many decision trees do we construct in total?
10. For the customer churn prediction task, what is the best choice of k [1-10] for the k-nearest neighbor algorithm in the 10-fold cross-validation scenario?
11. For MNIST dataset, what was the accuracy of the decision tree [max depth = 5] and K-nearest neighbor [K = 10]? What are the best parameters and accuracy for the holdout dataset for decision trees when we used GridSearchCV with 5-fold cross-validation?

Problem 3 [20 points]

Files for this problem are under the Experiment 3 folder.

Datasets for experimentation: spam.csv.

Jupyter notebook: Naive Bayes Spam.ipynb.

The dataset contains 5,574 messages tagged according to ham (legitimate) or spam. In this experiment, we will learn about text features, how to convert them into matrix form, and the Naive Bayes algorithm.

Answer the following questions:

1. How many records are there? What's the distribution of the “label” class? Is it skewed?
2. How many unique SMS messages are there in the dataset? What is the SMS message that occurs most frequently, and what is its frequency?
3. What is the maximum and minimum length of SMS messages present in the dataset? Plot the histogram of the length of SMS messages with bin sizes 5,10,20,50,100,200. What do you perceive after examining the plots?
4. Plot the histogram of the length of SMS messages for each label separately with bin sizes 5,10,20,50 i.e., histogram of the length of all ham SMS messages and histogram of the length of all spam SMS messages. What do you perceive after examining the plots?
5. In the Bag of words approach, we convert all strings into lower cases. Why did we do that, and why is it important? Can we convert all strings into the upper case and still fulfill our original goal?
6. What does CountVectorizer achieve? What will happen if we set stop words = "english"? Give five examples of stop-words in English.
7. Given a dataset, how do we generate a document-term matrix? Do we first generate a document-term matrix and then separate the matrix into train/test, or first separate the data into train/test and then generate a document-term matrix based on the training dataset and afterward generate a matrix for the test set? Explain your reasoning.
8. Using the bag of words approach, convert documents = ['Hi, how are you?', 'Win money, win from home. Call now.', 'Hi., Call you now or tomorrow?'] to its document-term matrix.
9. How many features are created while making a document-term matrix for the SMS dataset? Can you think of a method to reduce the number of features? List the pros and cons of the method.
10. For our input dataset, which Naive Bayes model should we use, Gaussian Naive Bayes or Multinomial Naive Bayes? Explain your reasoning. Report accuracy, precision, recall, and F1 score for the spam class after applying the Naive Bayes algorithm.

Problem 4 [40 points]

Files for this problem are under the Experiment 4 folder.

In this assignment, we provide three real-world datasets for classification:

Iris dataset (https://archive.ics.uci.edu/ml/datasets/Iris),

Thyroid dataset (https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease), and

Diabetes dataset (https://www.kaggle.com/uciml/pima-indians-diabetes-database).

Also, we give three Jupyter notebooks, one for each dataset, in which we have applied the k-nearest neighbor, decision tree, and naïve Bayes Algorithm without any parameter tuning.

Write a report (no longer than a page) for each dataset giving your analysis of the dataset, your observations, and comments. You can be as innovative as you want. Minimally, the report should include a brief description of the dataset, the number of observations, missing values or not, the testing strategy deployed, the classification accuracy of algorithms, intuition developed by running the notebooks, etc.

[Optional] Problem 5

For the three datasets that we have provided above, we used out-of-the-box classifiers for classification. Can you finetune the decision tree and kNN classifiers for the dataset using GridSearchCV? Report the finetuned value for the decision tree (max depth and max features) and the k-nearest neighbor (value of K).